



Multivariada I

- *Juliano van Melis – jvmelis@gmail.com*
- *Profa. MSc. Edmila Montezani*
- *edmila@gmail.com*



EXERCÍCIOS



Estudo de Caso: Salários MBA

- ❑ Perguntas que devem ser respondidas:
 1. Quanto os formandos podem esperar de salário após a formatura?
 2. Existem variáveis que tenham um efeito importante no valor do salário dos formandos (por exemplo: idade, gênero, quartil, língua mãe, experiência)?
 3. As informações fornecidas pelos programas de MBA são factíveis?

Estudo de Caso: Salários MBA

- A base de dados é apresentada a seguir. (mba.xlsx)
(várias linhas estão ocultas para facilitar a visualização).

age	sex	gmat_tot	gmat_qpc	gmat_vpc	gmat_tpc	s_avg	f_avg	quarter	work_yrs	frstlang	salary	satis
23	2	620	77	87	87	3,4	3	1	2	1	0	7
24	1	610	90	71	87	3,5	4	1	2	1	0	6
24	1	670	99	78	95	3,3	3,25	1	2	1	0	6
24	1	570	56	81	75	3,3	2,67	1	1	1	0	7
24	2	710	93	98	98	3,6	3,75	1	2	1	999	5
24	1	640	82	89	91	3,9	3,75	1	2	1	0	6
25	1	610	89	74	87	3,4	3,5	1	2	1	0	5
25	2	650	88	89	92	3,3	3,75	1	2	1	0	6
25	1	540	79	45	65	2,6	2,5	4	3	1	115000	5
26	1	550	72	58	69	2,6	2,75	4	3	1	126710	6
40	2	500	60	45	51	2,5	2,75	4	15	2	220000	6

- A resolução apresentada a seguir é baseada no software R.



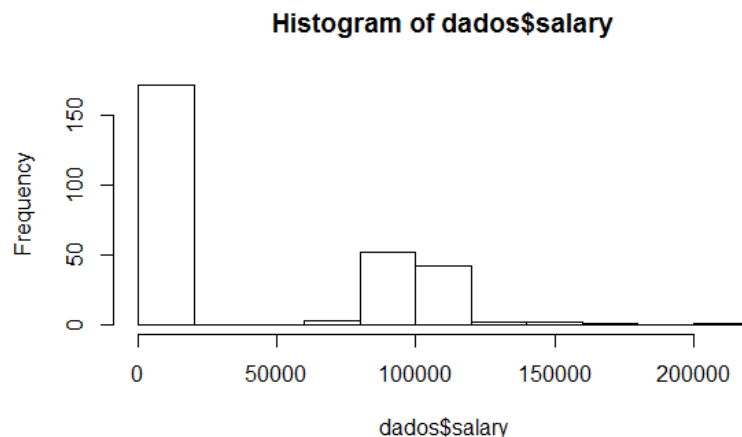
Estudo de Caso: Salários MBA

- ❑ 1º abra o RStudio. Na linha de comando digite:

```
dados=read.csv("C:/Users/Edmila/Desktop/Facu/Mack/Aulas/Arquivos aula 5/mba.csv", header = TRUE, sep = ";", dec=",")
```

Estudo de Caso: Salários MBA

- 1a) Quanto os formandos podem esperar de salário após a formatura?
- Primeiro deve-se determinar qual o salário médio dos estudantes após a formatura. A utilização dos dados relativos aos 274 alunos é uma estratégia razoável? Para ter uma idéia dos dados e um sumário estatístico no R podemos fazer:
 - `dados`
 - `names(dados)` #mostra os nomes das colunas da planilha mba.csv
 - `dados$salary` #mostra somente os dados de salário
 - `mean(dados$salary)` #média dos salários
 - `median(dados$salary)` #mediana dos salários
 - `hist(dados$salary)` #histograma dos salários



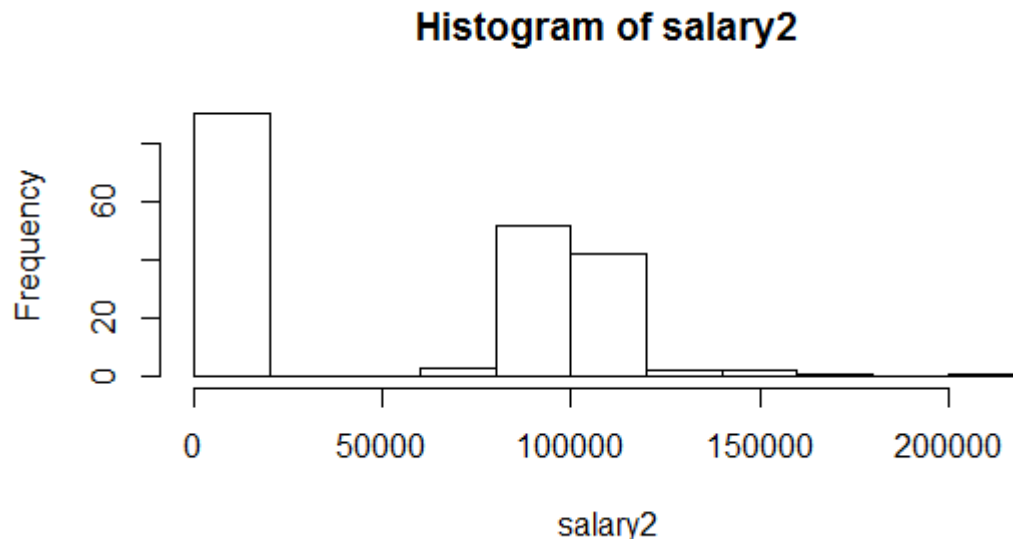
Estudo de Caso: Salários MBA

- ❑ O primeiro problema com a abordagem anterior é que não "limpamos" o dataset *salary* para fazer o cálculo da média. Pela descrição dos dados, vemos que 998 e 999 não representam um valor de salário, mas sim a ausência ou não fornecimento de infos (BUG do R....)
- ❑ Sendo assim, vamos criar uma variável *salary2* a qual terá apenas os valores de salário reais. A partir dela iremos calcular a média, mediana e histograma dos valores. Para isto fazemos:

```
salary2 = dados$salary[ (dados$salary != 999) & (dados$salary != 998) ]
```
- ❑ *dados\$salary != 999* significa valores de *dados\$salary* que sejam diferentes de 999. Isto vai gerar uma lista de FALSE e TRUE
- ❑ *&* é um operador lógico AND. A lista de FALSE e TRUE será gerada a partir dos valores de *salary* que seja diferentes de 999 E 998.

Estudo de Caso: Salários MBA

- Calculamos agora a média, mediana e histograma para *salary2*.
 - `mean(salary2)`
 - `median(salary2)`
 - `hist(salary2)`
- Obtivemos os valores de \$54.985,32 para a média e \$85.000 para a mediana. Além do histograma mostrado abaixo.



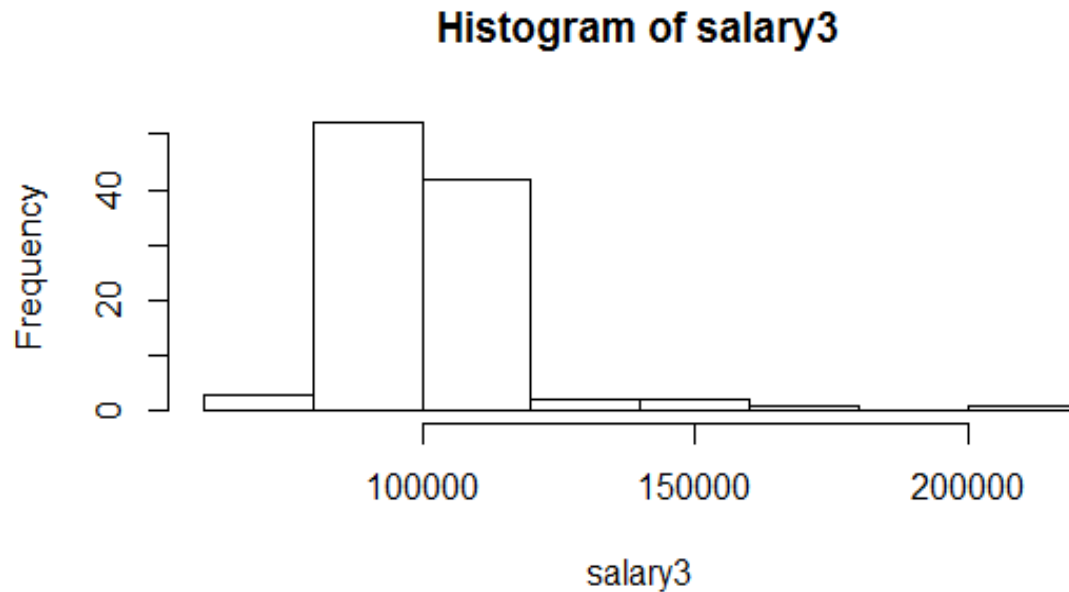


Estudo de Caso: Salários MBA

- ❑ O próximo ponto a ser levado em consideração é a grande quantidade de valores iguais a zero. Podemos obter o número de pontos totais, iguais e diferentes de zero fazendo:
 - `length(salary2)`
 - `sum(salary2==0)`
 - `sum(salary2>0)`
- ❑ Obtemos respectivamente 193, 90 e 103.
- ❑ Vamos agora criar um outro vetor de salários no qual estarão presentes apenas os salários maiores que 0 e diferentes de 999 e 998. Partindo de *salary2* fazemos:
 - `salary3 = salary2[salary2>0]`
 - `mean(salary3); median(salary3); hist(salary3);`

Estudo de Caso: Salários MBA

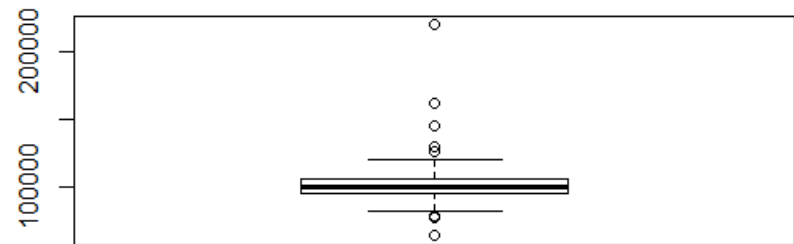
- Obtemos como resultados, para a média: \$103.030,7; para a mediana \$100.000 e o histograma:



- A maior similaridade da média e da mediana, além do próprio histograma garantem agora uma maior simetria, a qual por sua vez, pode ser interpretada como um sinal de melhor qualidade no *dataset* 😊

Estudo de Caso: Salários MBA

- ❑ Outra forma de obter o resumo dos dados seria:
 - `summary(salary3)`
 - Min. 1st Qu. Median Mean 3rd Qu. Max.
 - 64000 95000 100000 103000 106000 220000
- ❑ Ou poderíamos analisar os dados de forma visual através de um boxplot:
 - `boxplot(salary3)`

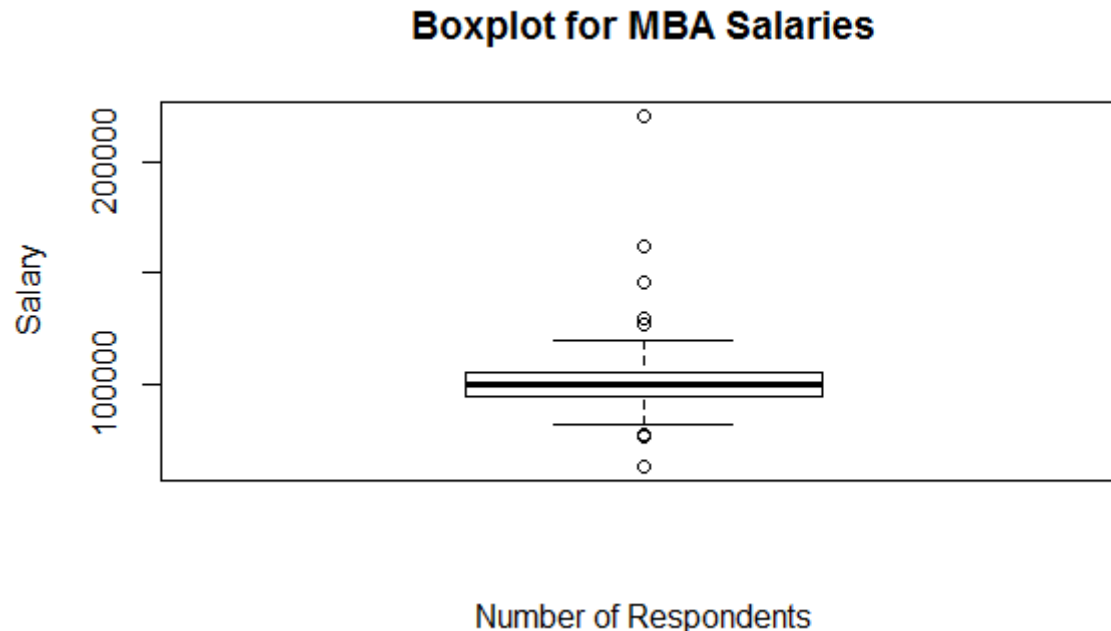


- ❑ No próximo slide veremos como melhorar o boxplot.

Estudo de Caso: Salários MBA

Opções para boxplot em ordem crescente de detalhes

- `boxplot(salary3, ylab="Salary")`
- `boxplot(salary3, ylab="Salary", xlab="Number of Respondents")`
- `boxplot(salary3, ylab="Salary", xlab="Number of Respondents", main="Boxplot for MBA Salaries")`

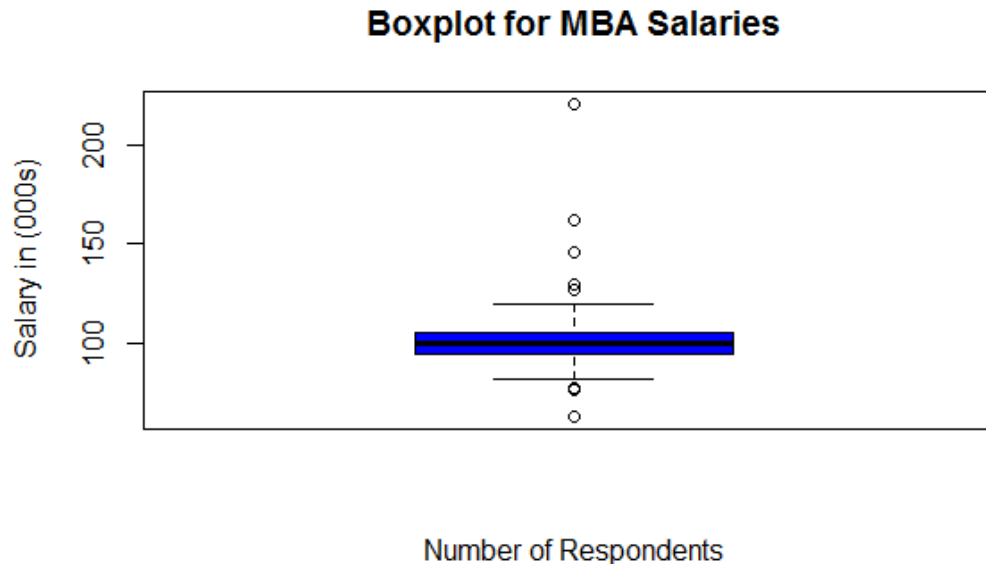


Estudo de Caso: Salários MBA

Opções para boxplot em ordem crescente de detalhes

.....

- `boxplot(salary3, ylab="Salary", xlab="Number of Respondents", main="Boxplot for MBA Salaries", col="blue")`
- `boxplot(salary3/1000, ylab="Salary in (000s)", xlab="Number of Respondents", main="Boxplot for MBA Salaries", col="blue")`



Estudo de Caso: Salários MBA

- Para gerar uma tabela de frequências e de frequências cumulativas de *salary3*:

■ Salary3

```
> salary3
[1] 85000 85000 86000 88000 92000 93000 95000 95000 95000 96000 96000
[12] 100000 100000 100000 105000 105000 105000 105000 105000 105000 106000 106000
[23] 107500 108000 110000 112000 115000 115000 118000 120000 120000 120000 120000
[34] 146000 162000 82000 92000 93000 95000 95000 96000 96500 98000 98000
[45] 98000 99000 100000 100000 101000 103000 104000 105000 105000 105000 107000
[56] 112000 115000 115000 130000 145800 78256 88500 90000 90000 93000 95000
[67] 97000 97000 98000 98000 98000 98000 98000 98000 100000 100000 101000
[78] 101100 102500 105000 106000 107300 108000 112000 64000 77000 85000 85000
[89] 86000 90000 92000 95000 96000 98000 100000 100000 100400 101600 104000
[100] 105000 115000 126710 220000
```

■ Table(salary3)

```
> table(salary3)
salary3
64000 77000 78256 82000 85000 86000 88000 88500 90000 92000 93000 95000
1 1 1 1 4 2 1 1 3 3 3 7
96000 96500 97000 98000 99000 100000 100400 101000 101100 101600 102500 103000
4 1 2 10 1 9 1 2 1 1 1 1
104000 105000 106000 107000 107300 107500 108000 110000 112000 115000 118000 120000
2 11 3 1 1 1 2 1 3 5 1 4
126710 130000 145800 146000 162000 220000
1 1 1 1 1 1
```



Estudo de Caso: Salários MBA

- ❑ Para gerar uma tabela de frequências e de frequências cumulativas de *salary3*:

... (continuação)

- `as.data.frame(table(salary3))` #visualização em colunas
- `salary4 = as.data.frame(table(salary3))`
- `salary4[,2]` #dados da segunda coluna
- `salary4$CumFreq = cumsum(salary4[,2])` #soma acumulada
- `salary4` #visualiza os dados em coluna com a adição da coluna com soma acumulada
- `names(salary4)` # nomes das colunas: [1] "salary3" "Freq" "CumFreq"
- `salary4$PercFreq= salary4[,2]/sum(salary4[,2])` #% da coluna Freq
- `salary4$PercCumFreq= cumsum(salary4[,3])/sum(salary4[,2])`
- #% da coluna Freq Acumulada
- `salary4` #Apresenta a tabela de frequências



Estudo de Caso: Salários MBA

- ❑ Para obter os valores numéricos dos fatores de classificação de *salary4* podemos fazer:
 - `salary4[,1]` #Fatores não numéricos e sua descrição
 - `levels(salary4[,1])` #Apenas os fatores não numéricos
 - `as.numeric(levels(salary4[,1]))` #Fatores numéricos como números
 - `niveis = as.numeric(levels(salary4[,1]))` #tabela como números

- ❑ Em seguida para inserir os níveis numéricos na segunda coluna do data frame (criando outro data frame agora com o nome *salary5*):
 - `salary5 = data.frame(salary4[,1], niveis, salary4[,2], salary4[,3], salary4[,4], salary4[,5])`

- ❑ No entanto isto requer uma grande digitação de dados repetidos. Podemos abreviar a digitação fazendo a sequência a seguir (próximo slide):

Estudo de Caso: Salários MBA

❑ Inserção de dados (colunas) em data frame (entre a 1ª e a 2ª coluna):

- 2:5
- salary4
- salary4[,2:5] #tira a informação de Salários
- niveis = as.numeric(levels(salary4[,1]))
- data.frame(salary4[,1],niveis,salary4[,2:5])
- #encontra a tabela de frequencias da tabela “niveis”
- salary5= data.frame(salary4[,1],niveis,salary4[,2:5])
- names(salary5)[1] = "SalaryLevels"
- names(salary5)[2] = "Níveis"
- salary5 →
- names(salary5)

```
> salary5
  SalaryLevels Níveis Freq CumFreq   PercFreq PercCumFreq
1         64000  64000   1      1 0.009708738 0.009708738
2         77000  77000   1      2 0.009708738 0.029126214
3         78256  78256   1      3 0.009708738 0.058252427
```

```
> names(salary5)
[1] "SalaryLevels" "Níveis"       "Freq"         "CumFreq"      "PercFreq"
[6] "PercCumFreq"
```

Estudo de Caso: Salários MBA

□ Status Atual:

- `summary(salary3)`

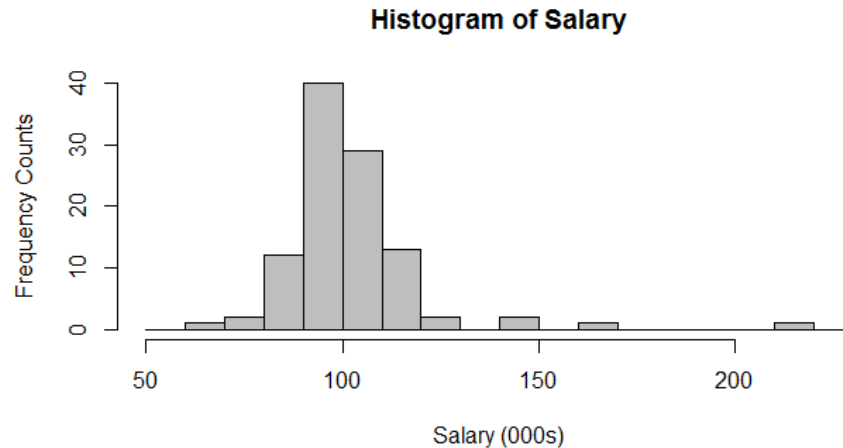
- Média: \$103.000

- Mediana: \$100.000

```
> summary(salary3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 64000   95000  100000   103000  106000  220000
```

○ Histograma de Salários

- `hist(salary3/1000, breaks=seq(from=50, to = 230, by=10), col="gray", xlab="Salary (000s)", ylab="Frequency Counts", main="Histogram of Salary")`

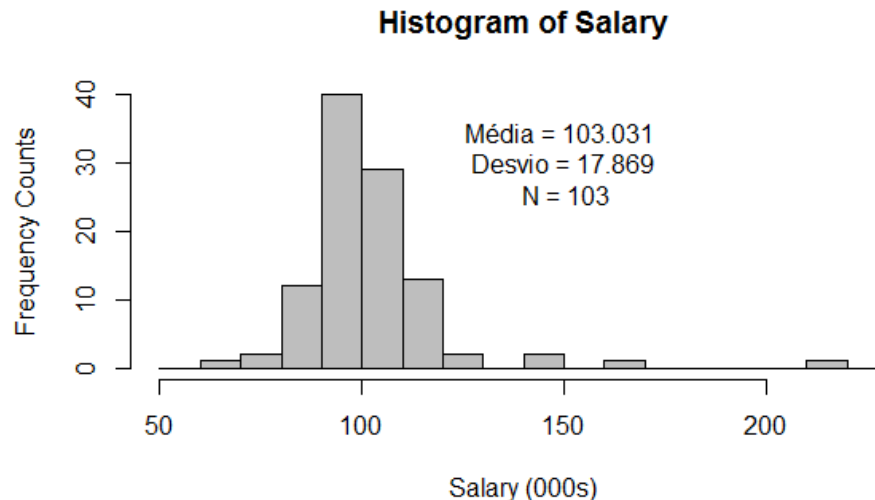


□ Histograma de Salários com texto informando média, desvio e contagem..... (Próximo slide)

Estudo de Caso: Salários MBA

- Histograma de Salários com texto informando média, desvio e contagem

```
plot.new();  
  
hist(salary3/1000, breaks=seq(from=50, to = 230, by=10),  
     col="gray", xlab="Salary (000s)", ylab="Frequency Counts",  
     main="Histogram of Salary");  
  
text(150,30,paste("Média =", format(round(mean(salary3),0),  
                                decimal.mark = ",", big.mark = "."),  
                "\n Desvio =", format(round(sd(salary3),0),  
                                decimal.mark = ",", big.mark = "."),  
                "\n N =", length(salary3)));
```



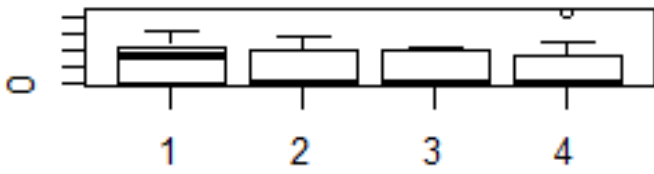
Estudo de Caso: MBA Salaries

2a) Existem variáveis que afetam o valor esperado do salário inicial de um recém formado?

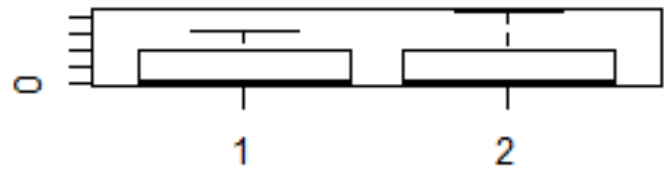
- Podem ser analisados: notas, gênero, língua mãe e experiência. Vamos analisar o efeito através de boxplots que serão divididos pelas categorias dos quatro fatores, através do código abaixo (resultado a seguir):
 - `layout(matrix(c(1,2,3,4),2,2,byrow=TRUE))`
 - `boxplot(dados$salary~dados$quarter, main="Salary by Quarter")`
 - `boxplot(dados$salary~dados$sex, main="Salary by gender")`
 - `boxplot(dados$salary~dados$frstlang, main="Salary by mother tongue")`
 - `boxplot(dados$salary~dados$work_yrs, main="Salary by work experience")`
 - `layout(1,1,1)`

Estudo de Caso: MBA Salaries

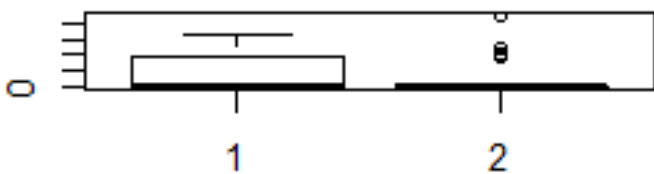
Salary by Quarter



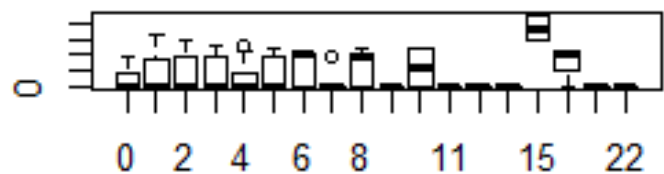
Salary by gender



Salary by mother tongue



Salary by work experience



Estudo de Caso: MBA Salaries

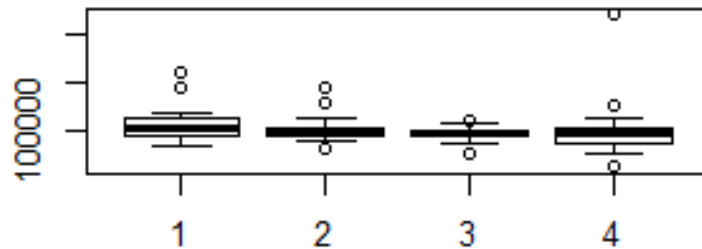
As comparações do slide anterior foram feitas no entanto considerando-se todos os elementos do dataset *salary*.

Devemos lembrar que nossa análise desconta os elementos de *salary* com valores iguais a 0, 999 ou 0,998. Para criar gráficos com estes dados podemos fazer:

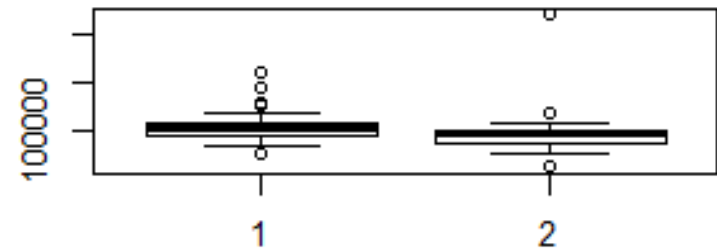
- `layout(matrix(c(1,2,3,4),2,2,byrow=TRUE))`
- `attach(dados)`
- `salary2 = salary[salary!=0 & salary!=998 & salary!=999]`
- `quarter2 = quarter[salary!=0 & salary!=998 & salary!=999]`
- `sex2 = sex[salary!=0 & salary!=998 & salary!=999]`
- `frstlang2 = frstlang[salary!=0 & salary!=998 & salary!=999]`
- `work_yrs2 = work_yrs[salary!=0 & salary!=998 & salary!=999]`
- `boxplot(salary2~quarter2, main="Salary by Quarter")`
- `boxplot(salary2~sex2, main="Salary by gender")`
- `boxplot(salary2~frstlang2, main="Salary by mother tongue")`
- `boxplot(salary2~work_yrs2, main="Salary by work experience")`
- `detach(dados)`
- `layout(1,1,1)`

Estudo de Caso: MBA Salaries

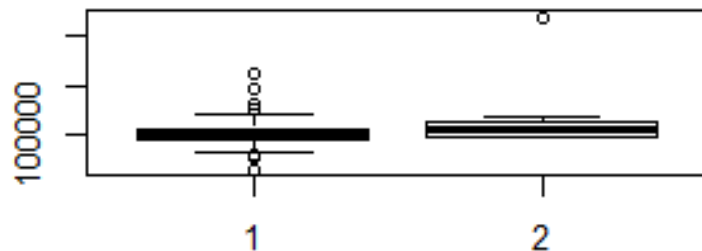
Salary by Quarter



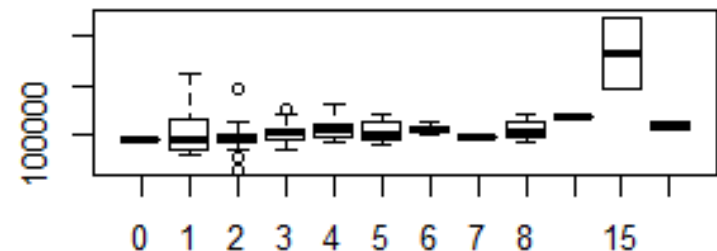
Salary by gender



Salary by mother tongue



Salary by work experience



Estudo de Caso: MBA Salaries

A comparação também pode ser executada através de cálculos diretos nos datasets. No caso abaixo, as medidas estatísticas foram calculadas no R com os comandos:

- `summary(salary2[quarter2==1]);sd(salary2[quarter2==1]); length(salary2[quarter2==1])`
- `summary(salary2[quarter2==2]);sd(salary2[quarter2==2]); length(salary2[quarter2==2])`
- `summary(salary2[quarter2==3]);sd(salary2[quarter2==3]); length(salary2[quarter2==3])`
- `summary(salary2[quarter2==4]);sd(salary2[quarter2==4]); length(salary2[quarter2==4])`

Table 5- Comparação de Quartis				
	Q1	Q2	Q3	Q4
Média	\$106.328	\$103.612	\$98.319	\$102.142
Mediana	\$105.000	\$100.000	\$98.000	\$98.000
Mínimo	\$85.000	\$82.000	\$78.526	\$64.000
Máximo	\$162.000	\$145.800	\$112.000	\$190.000
Intervalo	\$77.000	\$63.800	\$33.744	\$222.000
Desvio	\$15.838	\$12.818	\$7.175	\$31.600
Tam. Am.	35	25	24	19

Comentários: 1) Aumento da média no 4º quartil -> decorre de um elemento *outlier* (220k), referente ao salário de aluno que retornou ao seu país de origem para assumir os negócios da família + pequeno tamanho da amostra neste quartil.

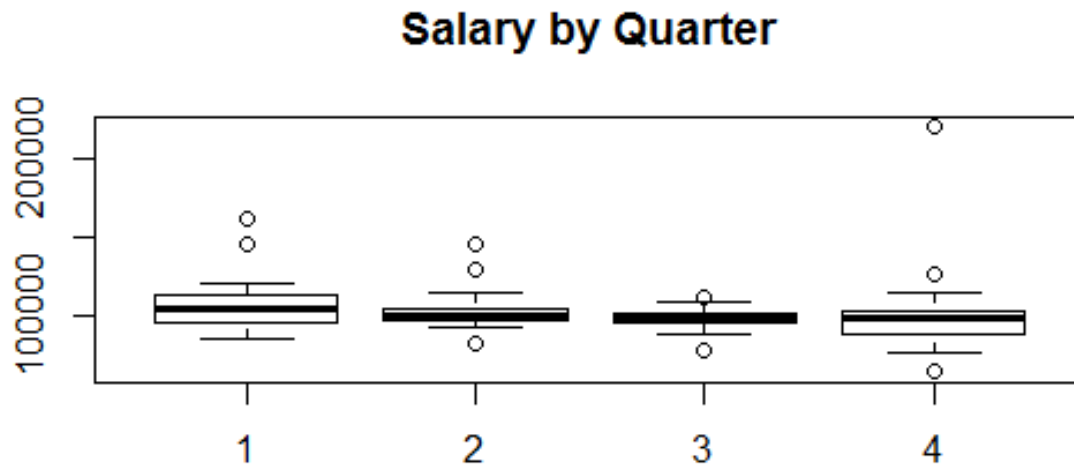
Comentários: 2) Melhor quartil para previsões: 3º por conta do menor desvio

Estudo de Caso: MBA Salaries – Teste ANOVA – Influência da Nota no Salário

- ❑ Vamos agora realizar um teste ANOVA, para avaliar se existe diferença na média dos salários quando segmentados por quartil de nota. Observe a forma como as tabelas de dados foram geradas a partir do data frame *dados*
 - `attach(dados)`
 - `names(dados)` # nomes das colunas
 - `dados2 = dados[salary!=0 & salary!=999 & salary!=998,]` # retirar as sujeiras
 - `detach(dados)`
 - `attach(dados2)`
 - `names(dados2)`
 - `oneway.test(salary~quarter, data=dados, var.equal=TRUE)`
- ❑ Obtemos como resultado:
One-way analysis of means data: salary and quarter $F = 2.993$, num df = 3, denom df = 270, p-value = 0.03136
- ❑ O p-value de *0.031* aponta para a não existência de diferença entre as médias salariais por quartil.

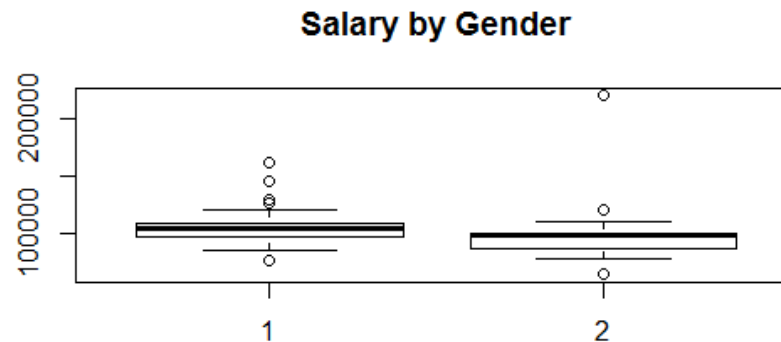
Estudo de Caso: MBA Salaries – Teste ANOVA – Influência da Nota no Salário

- ❑ Podemos também gerar rapidamente um boxplot de *salary* separado por *quarter* fazendo o seguinte:
 - `boxplot(salary~quarter, main="Salary by Quarter")`
- ❑ O teste ANOVA é visualmente confirmado pelo boxplot, pois os diagramas não mostram diferença significativa entre as média



Estudo de Caso: MBA Salaries – Teste ANOVA – Influência do gênero no Salário

- ❑ Gerando o boxplot:
 - `boxplot(salary~sex, main="Salary by Gender")`



- ❑ Não parece haver uma grande diferença, porém calculamos podemos calcular o ANOVA para confirmar
 - `oneway.test(salary~sex, data=dados2, var.equal=TRUE)`
- ❑ Obtemos como resultado:
 - One-way analysis of means data: salary and sex $F = 2.87$, num df = 1, denom df = 101, p-value = 0.0932

Teste ANOVA – Influência do gênero no Salário

- ❑ **No slide anterior, executamos o ANOVA sob a pressuposição de variâncias equivalentes. Vamos testar esta pressuposição.**
 - `var.test(salary~sex)`
- ❑ **Obtemos como resultado:**
 - F test to compare two variances data: salary by sex F = 0.30486, num df = 71, denom df = 30, p-value = 4.241e-05 alternative hypothesis: true ratio of variances is not equal to 1 95 percent confidence interval: 0.1589648 0.5414794 sample estimates: ratio of variances 0.3048572
 - **Executamos então um teste t com dados não emparelhados, variâncias distintas.**
 - `t.test(salary~sex, paired=FALSE, var.equal=FALSE)`
- ❑ **Neste caso obtemos como resultado:**
 - Welch Two Sample t-test data: salary by sex t = 1.36, df = 38.1, p-value = 0.1809 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -3129 16022 sample estimates: mean in group 1 mean in group 2
104971 98524
- ❑ **Poderíamos também ter executado o teste ANOVA sob a pressuposição de variâncias distintas fazendo:**
 - `oneway.test(salary~sex, data=dados2, var.equal=FALSE)`
- ❑ **Obtendo como resultado:**
 - One-way analysis of means (not assuming equal variances) data: salary and sex F = 1.86, num df = 1.0, denom df = 38.1, p-value = 0.1809

ANOVA 2 way – Influência do quartil e do gênero no salário

- ❑ Para isto precisamos utilizar outra função *aov()*, as regras de modelagem multivariada do R e a função *summary*.
 - `result=aov(salary~quarter*sex, data=dados2)`
 - `summary(result)`

```
> summary(result)
              Df    Sum Sq   Mean Sq F value Pr(>F)
quarter       1  5.376e+08  5.376e+08   1.809  0.1817
sex           1  9.310e+08  9.310e+08   3.133  0.0798 .
quarter:sex    1  1.680e+09  1.680e+09   5.653  0.0193 *
Residuals    99  2.942e+10  2.972e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

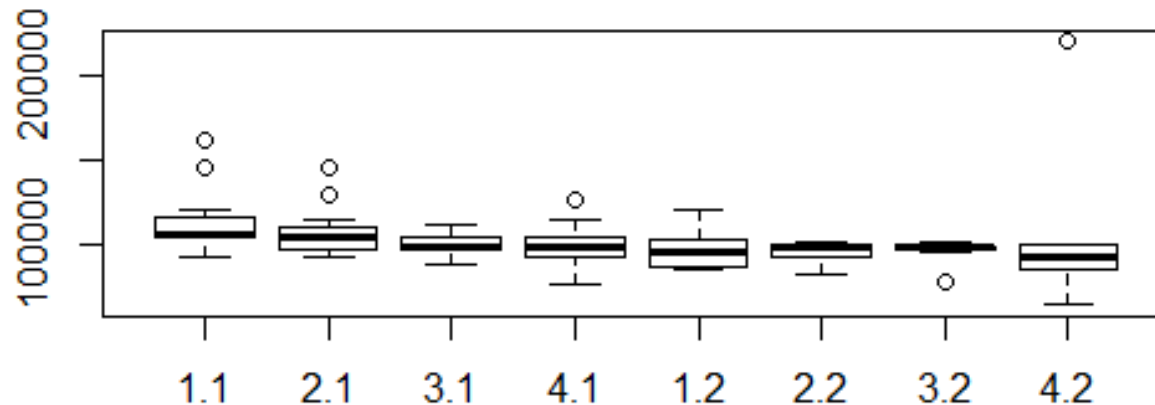
Fazendo *quarter*sex* estamos testando possíveis interações entre os quartis e o gênero. Poderíamos ter feito *quarter+sex* o que tornaria o modelo linear (aditivo, sem interações).

No entanto a análise do resultado de *summary* mostra que os *p-values* foram: 0,182; 0,080 e 0,019 para *quarter*, *sex* e *interações (quarter x sex)* respectivamente.

Isto mostra que ocorre uma interação (evento específico) de quartil com gênero.

ANOVA 2 way – Influência do quartil e do gênero no salário

- Fazendo o boxplot:
 - `boxplot(salary~quarter*sex, data=dados2)`
- Obtemos: isto mostra que no 4º quartil
 - gênero = 2 (fem) existe um salário discrepante





Estudo de Caso: MBA Salaries – Efeito da Língua Mãe

- Neste caso não é viável realizar este teste pois apenas 7 dos respondentes afirmaram não ser o inglês sua língua mãe.
- Isto pode ser confirmado fazendo-se:
 - `sum(frstlang==2)`
 - 7

MBA Salaries – Efeito da Experiência, Nível de Satisfação e Notas

- ❑ Neste caso vamos utilizar uma tabela de correlações para observar se existe uma relação entre o salário e os fatores experiência, notas e nível de satisfação.
- ❑ Para tal precisaremos dos seguintes comandos
 - `names(dados2)`
 - `mbacor = data.frame(gmat_tot, salary, s_avg, work_yrs, satis)`
- ❑ Para obter o no.de observações
 - `length(mbacor[,1])`
- ❑ Para obter a matriz de correlações, utilizasse a função `cor()` a qual aceita data frame como input
 - `cor(mbacor)`

```
> cor(mbacor)
```

	gmat_tot	salary	s_avg	work_yrs	satis
gmat_tot	1.00000000	-0.09067141	0.1719887	-0.12280018	0.06474206
salary	-0.09067141	1.00000000	0.1017317	0.45466634	-0.04005060
s_avg	0.17198874	0.10173175	1.0000000	0.16328236	-0.14356557
work_yrs	-0.12280018	0.45466634	0.1632824	1.00000000	0.06299926
satis	0.06474206	-0.04005060	-0.1435656	0.06299926	1.00000000

MBA Salaries – Efeito da Experiência, Nível de Satisfação e Notas

- Caso queira-se saber também o p-valor de todas as correlações encontradas pode-se utilizar a função `rcorr()` do pacote *Hmisc*. Esta função no entanto aceita apenas matrizes como input. Sendo assim deve-se utilizar:

- `library(Hmisc)`
- `rcorr(as.matrix(mbacor))`

```
> rcorr(as.matrix(mbacor))
```

	gmat_tot	salary	s_avg	work_yrs	satis
gmat_tot	1.00	-0.09	0.17	-0.12	0.06
salary	-0.09	1.00	0.10	0.45	-0.04
s_avg	0.17	0.10	1.00	0.16	-0.14
work_yrs	-0.12	0.45	0.16	1.00	0.06
satis	0.06	-0.04	-0.14	0.06	1.00

n= 103

P

	gmat_tot	salary	s_avg	work_yrs	satis
gmat_tot		0.3624	0.0824	0.2165	0.5159
salary	0.3624		0.3065	0.0000	0.6879
s_avg	0.0824	0.3065		0.0994	0.1480
work_yrs	0.2165	0.0000	0.0994		0.5273
satis	0.5159	0.6879	0.1480	0.5273	

- No caso da função padrão do R `cor.test` a mesma aceita apenas pares de vetores. Sendo assim seria necessário utilizar combinações como:

- `cor.test(mbacor[,1], mbacor[,2])`

```
> cor.test(mbacor[,1], mbacor[,2])
```

- Resultados são apresentados a seguir:

```
Pearson's product-moment correlation
```

data: mbacor[, 1] and mbacor[, 2]
t = -0.91501, df = 101, p-value = 0.3624
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.2792952 0.1046903
sample estimates:
cor
-0.09067141



MBA Salaries – Análise de Correlações e Modelo de Regressão

- ❑ A única correlação significativa ocorre entre salário e anos de experiência ($r = .455$, $p < .001$).
- ❑ A partir das variáveis pede-se agora construir um modelo de regressão de modo a prever o valor do salário inicial. O modelo inicial (o qual inclui gênero, GMAT score, notas, anos de experiência, língua mãe e idade) é apresentado a seguir.
- ❑ O modelo foi desenvolvido através dos comandos:
 - `names(dados2)`
 - `lm(salary~gmat_tot+s_avg+age+sex+frstlang+age+work_yrs)->a`
 - `summary(a)`

MBA Salaries – Análise de Correlações e Modelo de Regressão

```
> summary(a)
```

```
Call:
```

```
lm(formula = salary ~ gmat_tot + s_avg + age + sex + frstlang +  
    age + work_yrs)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-31300	-8779	-2326	6045	80330

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	52225.51	31350.22	1.666	0.099 .
gmat_tot	-16.08	31.58	-0.509	0.612
s_avg	3440.28	4332.59	0.794	0.429
age	1570.27	1110.24	1.414	0.160
sex	-5044.31	3474.99	-1.452	0.150
frstlang	10755.93	7041.44	1.528	0.130
work_yrs	843.48	1138.46	0.741	0.461


```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 15610 on 96 degrees of freedom
```

```
Multiple R-squared:  0.2821,    Adjusted R-squared:  0.2372
```

```
F-statistic: 6.287 on 6 and 96 DF,  p-value: 1.333e-05
```



Regressão no Excel 😊
(Pode ficar feliz, Thiago!!)

Outro Exemplo - Manutenção do caminhão

Uma agroindústria quer saber o custo de manutenção de seus caminhões durante o corrente ano, para tanto foram coletadas informações de quilometragem e tempo do caminhão. A tabela abaixo nos mostra esses valores.

Custo de Manutenção	Quilometragem (x1000)	Tempo do caminhão (em anos)
832	6	8
73	7	7
647	9	6
553	11	5
467	13	4
373	15	3
283	17	2
189	18	1
96	19	0

Resolução

Nesse caso será feito diretamente análise sem plotar o gráfico.

O procedimento no software Excel é: Ferramenta -> Análise de Dados -> Regressão.

No campo Intervalo X de Entrada deve ser preenchida com a faixa de valores das variáveis independentes, que nesse caso são a quilometragem e o tempo do caminhão.



Regressão Não Linear



Regressão Não-Linear

Nem sempre a relação entre a variável independente (X) e a variável dependente (Y) possui uma relação linear, em certos casos essa relação é não-linear.

Nesses casos, pode-se através de mudanças de variáveis resolver o problema utilizando basicamente as equações já mencionadas nesse material.

Para efeito de demonstração da Regressão-Linear será utilizado o Excel através do seu recurso de Tendência, todavia conforme já mencionado, esse não dá informações estatísticas sobre o ajuste.

Vamos ver um exemplo.....

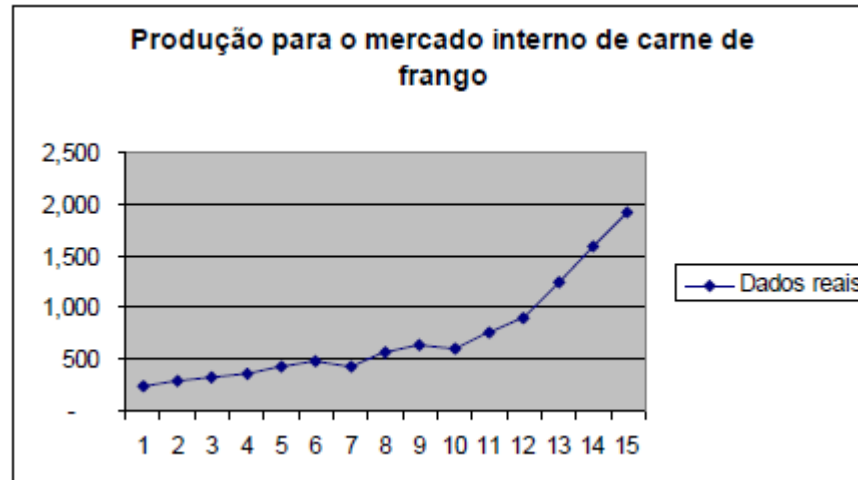
Exemplo: Série Temporal da Produção de Carne de Frango no Brasil ((1989--2003))

De acordo com a Associação Brasileira de Exportadora dos Produtores e Exportadores de Frango, ABEF, a produção brasileira de carne de frango (em mil toneladas) para o mercado interno e externo no período de 1989 a 2003 é dada pela tabela abaixo:

Ano	Mercado Interno	Exportação	Total
1989	1,811	244	2,055
1990	1,968	299	2,267
1991	2,200	322	2,522
1992	2,351	372	2,727
1993	2,710	433	3,143
1994	2,930	481	3,411
1995	3,617	429	4,050
1996	3,483	569	4,052
1997	3,812	649	4,461
1998	4,262	612	4,875
1999	4,755	771	5,526
2000	5,070	907	5,977
2001	5,486	1,249	6,736
2002	5,917	1,600	7,517
2003	5,921	1,922	7,843

Resolução

Nesse exemplo será avaliada somente a produção para o mercado externo, o gráfico que representa essa produção ao longo do ano pode ser visto logo abaixo.



Resolução

Pelo gráfico percebe-se uma tendência que a relação entre a produção de carne de frango (variável dependente, Y) e o tempo (variável independente, X) seja dado por uma equação linear. Para determinar essa equação será utilizado o software Excel.

No Excel será utilizada a ferramenta Regressão que é um módulo do Suplemento Análise de Dados.

Resolução

The screenshot shows the Microsoft Excel 2003 interface. The 'Ferramentas' (Tools) menu is open, displaying various options. The 'Análise de dados...' (Data Analysis...) option is highlighted. Below this, the 'Análise de dados' (Data Analysis) dialog box is open, showing a list of data analysis tools. The 'Regressão' (Regression) tool is selected in the list. The background spreadsheet contains data for years 1989 to 2003, with columns for 'Regiões', 'Mercado Interno', and 'Exportação'.

Microsoft Excel - Plan1

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Cell Run Ajuda

Verificar ortografia... F7
AutoSalvamento...
Compartilhar pasta de trabalho...
Proteger
Solver...
Suplementos...
Personalizar...
Opções...
Assistente
Análise de dados...
Atualizar vínculos de suplementos...

Análise de dados

Ferramentas de análise

- Análise de Fourier
- Histograma
- Média móvel
- Geração de número aleatório
- Ordem e percentil
- Regressão**
- Amostragem
- Teste-T: duas amostras em par para médias
- Teste-T: duas amostras presumindo variâncias equivalentes
- Teste-T: duas amostras presumindo variâncias diferentes

OK
Cancelar
Ajuda

	A	B	C
1	Regiões	Mercado Interno	Exportação
2			
3			
4	1989	1,811	244
5	1990	1,968	299
6	1991	2,200	322
7	1992	2,351	372
8	1993	2,710	433
9	1994	2,930	481
10	1995	3,617	429
11	1996	3,483	569
12	1997	3,812	649
13	1998	4,262	612
14	1999	4,755	771
15	2000	5,070	907
16	2001	5,486	1,249
17	2002	5,917	1,600
18	2003	5,921	1,922
19			
20			

Resolução

Acionando-se essa ferramenta, o passo seguinte será preencher a caixa de diálogo da Regressão conforme os dados.

Onde na opção Intervalo Y de Entrada deverá ser colocado o valor da variável dependente, e na opção Intervalo X de Entrada, deverá ser colocado os valores da variável independente.

Regressão

Entrada

Intervalo Y de entrada:

Intervalo X de entrada:

☐ Rótulos ☐ Constante é zero

☐ Nível de confiança: %

Opções de saída

☐ Intervalo de saída:

☒ Nova planilha:

☐ Nova pasta de trabalho

Resíduos

☐ Resíduos ☐ Plotar resíduos

☐ Resíduos padronizados ☒ Plotar ajuste de linha

Probabilidade normal

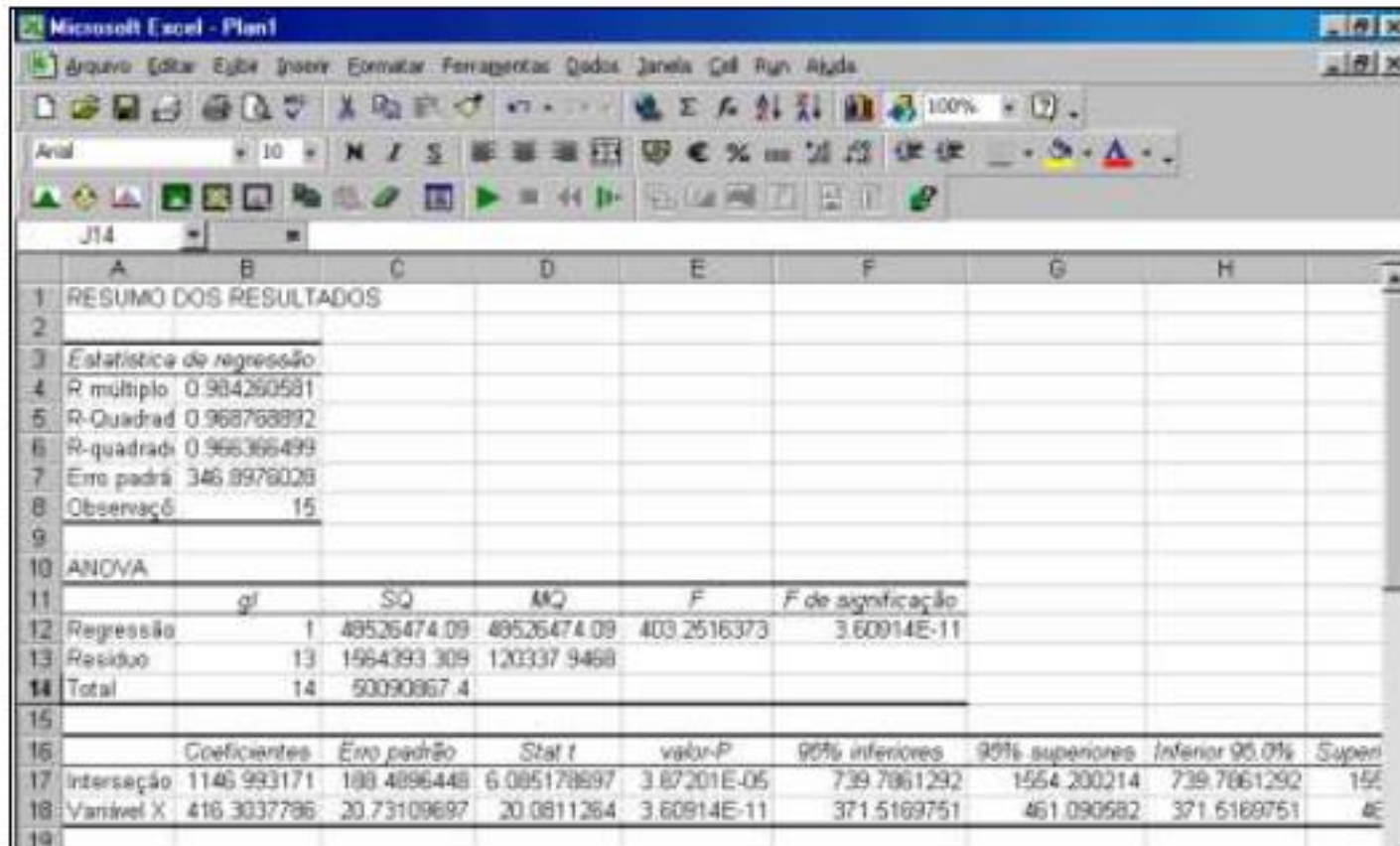
☐ Plotagem de probabilidade normal

OK
Cancelar
Ajuda

Resolução

Após o preenchimento das caixas de diálogo basta pressionar o botão de *Ok*, e o resultado aparecerá em uma nova planilha.

A figura abaixo mostra o resultado para o exemplo em questão.



The screenshot shows a Microsoft Excel window titled 'Plan1' with a menu bar (Arquivo, Editar, Exibir, Inserir, Formatar, Ferramentas, Dados, Janela, Graf, Ajuda) and a toolbar. The active sheet is 'J14'. The data is organized into two main sections: 'Estatística de regressão' and 'ANOVA'.

Estatística de regressão							
R múltiplo	0.984260581						
R-Quadrado	0.968768892						
R-quadrado	0.968366499						
Erro padrão	346.8978028						
Observações	15						

ANOVA					
	gl	SS	MS	F	F de significação
Regressão	1	49526474.09	49526474.09	403.2516373	3.60914E-11
Resíduo	13	1564393.309	120337.9468		
Total	14	50090867.4			

	Coefficientes	Erro padrão	Stat t	valor-P	90% inferiores	90% superiores	Inferior 95.0%	Superior 95.0%
Interseção	1146.993171	188.4896448	6.085178897	3.87201E-05	739.7861292	1554.200214	739.7861292	1554.200214
Variável X	416.3037796	20.73109697	20.0811264	3.60914E-11	371.5169751	461.090582	371.5169751	461.090582

Resolução

Dessa planilha se destacam os seguintes valores:

Na estatística padrão: $R\text{-quadrado} = 0.9687$

Na Anova:

$gl\ total = 14$

$F = 403.251$

E por fim:

Interseção: 1146,99

Variável X: 416,30

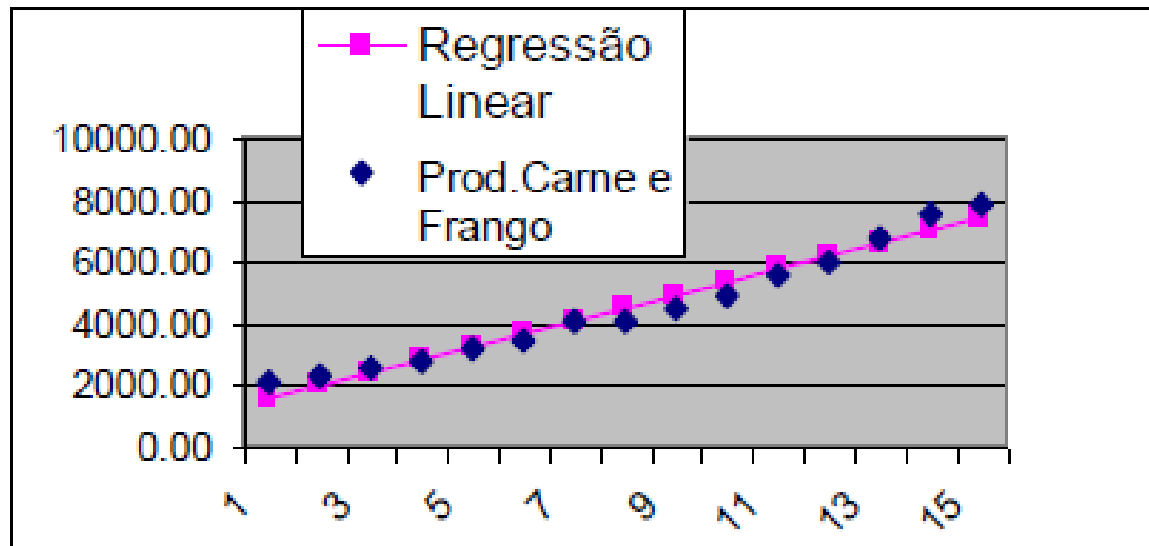
Assim a equação do modelo poderá ser escrita como:

$$\hat{Y} = 1146,99 + 416,30X_i$$

Resolução

Pode-se agora plotar os dados dos valores verdadeiros com os valores do modelo.

Também se pode fazer prognóstico para valores futuros. Por exemplo, para o ano de 2004 o modelo prevê uma produção de 7.807 toneladas de carne de frango.

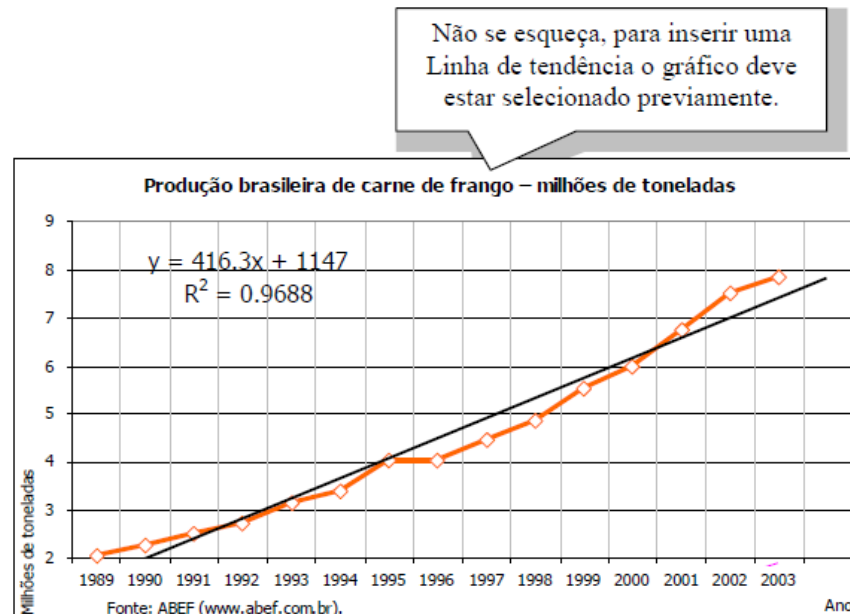


Resolução

Uma outra maneira de fazer essa análise, porém sem as mesmas informações seria utilizar o recurso de *Adicionar Linha de Tendência...* no Menu Gráfico da barra de menu do Excel.

Selecionado o modelo Linear, clica-se na aba Opções e marca-se as opções:

Exibir equação no gráfico e Exibir valor do R-quadrado no gráfico.



Resolução

Rebanho bovino brasileiro – efetivo por estado
(Mil cabeças)

Regiões	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
<i>Norte</i>	13,317	15,362	15,847	17,067	17,966	19,183	17,983	19,298	21,099	22,431	24,518	27,284	30,429
RO	1,719	2,826	2,774	3,286	3,470	3,928	3,937	4,331	5,104	5,442	5,664	6,605	8,040
AC	400	404	409	445	465	471	853	863	907	930	1,033	1,673	1,817
AM	637	648	640	689	747	806	734	771	809	826	843	864	895
RR	-	346	349	-	286	282	400	378	425	481	480	438	423
PA	6,182	6,626	6,990	7,435	7,539	8,058	6,751	7,539	8,337	8,863	10,271	11,047	12,191
AP	70	71	62	73	86	93	64	66	75	77	83	87	84
TO	4,309	4,441	4,624	5,139	5,374	5,544	5,243	5,351	5,442	5,813	6,142	6,571	6,979
<i>Nordeste</i>	26,190	26,669	26,912	22,527	22,825	23,174	23,882	23,831	21,981	21,875	22,567	23,414	23,891
MA	3,900	3,949	3,931	4,020	4,102	4,162	3,936	3,905	3,937	3,966	4,094	4,483	4,776
PI	1,974	2,046	2,029	1,982	2,054	2,135	1,730	1,737	1,751	1,756	1,779	1,792	1,804
CE	2,621	2,625	2,602	2,098	2,186	2,266	2,400	2,411	2,114	2,168	2,206	2,194	2,230
RN	956	966	930	566	646	722	935	941	793	755	804	788	839
PB	1,345	1,315	1,320	859	975	1,054	1,305	1,303	929	886	953	918	952
PE	1,966	1,952	1,923	1,271	1,349	1,362	1,954	1,682	1,470	1,420	1,516	1,673	1,753
AL	891	961	959	802	822	834	839	956	900	815	779	843	816
SE	1,030	1,047	1,058	908	815	797	946	946	918	937	880	866	863
BA	11,505	11,808	12,160	10,022	9,877	9,841	9,838	9,950	9,168	9,171	9,557	9,856	9,856
<i>Sudeste</i>	36,323	36,724	37,231	37,627	37,604	37,168	36,605	36,977	37,074	36,899	36,852	37,119	37,924
MG	20,472	20,764	21,066	21,034	20,707	20,146	20,148	20,378	20,501	20,082	19,975	20,219	20,559
ES	1,665	1,766	1,829	1,935	1,919	1,968	1,816	1,936	1,938	1,882	1,825	1,665	1,683
RJ	1,924	1,932	1,942	1,967	2,004	1,905	1,843	1,837	1,881	1,866	1,959	1,977	1,981
SP	12,263	12,262	12,394	12,690	12,974	13,148	12,798	12,827	12,753	13,069	13,092	13,258	13,701
<i>SUL</i>	25,326	25,272	25,451	25,727	26,429	26,641	26,421	26,683	26,600	26,190	26,298	26,784	27,537
PR	8,617	8,542	8,499	8,607	8,912	9,389	9,880	9,897	9,767	9,473	9,646	9,817	10,048
SC	2,994	3,057	3,047	3,017	2,960	2,993	3,098	3,087	3,090	3,053	3,051	3,096	3,118
RS	13,715	13,673	13,905	14,103	14,556	14,259	13,443	13,700	13,743	13,664	13,601	13,872	14,371
<i>Centro-Oeste</i>	45,946	48,109	48,788	52,186	53,420	55,061	53,398	54,627	56,402	57,227	59,641	61,787	65,567
MS	19,164	19,543	20,395	21,800	22,244	22,292	20,756	20,983	21,422	21,576	22,205	22,620	23,168
MT	9,041	9,891	10,138	11,682	12,654	14,154	15,573	16,338	16,752	17,243	18,925	19,922	22,184
GO	17,635	18,574	18,148	18,581	18,397	18,492	16,955	17,182	18,118	18,297	18,399	19,132	20,102
DF	106	102	107	124	124	123	115	123	110	110	112	113	113
Brasil	147,102	152,136	154,229	155,134	158,243	161,228	158,289	161,416	163,154	164,621	169,876	176,389	185,347

Fonte: IBGE – Pesquisa Pecuária Municipal (www.ibge.gov.br).



Obrigada!

Edmila Montezani
edmila@gmail.com